Eriksson, Katherine

Article

The promise of linked historical census data

NBER Reporter

Provided in Cooperation with:
National Bureau of Economic Research (NBER), Cambridge, Mass.

Suggested Citation: Eriksson, Katherine (2022) : The promise of linked historical census data, NBER Reporter, ISSN 0276-119X, National Bureau of Economic Research (NBER), Cambridge, MA, Iss. 2, pp. 7-9

This Version is available at:
https://hdl.handle.net/10419/265488

The Promise of Linked Historical Census Data

Katherine Eriksson

Individual records from the 1950 US Census were publicly released on April 1, 2022. Economic historians had been waiting for this day for 10 years. This data source, like the individual-level data from earlier censuses, makes it possible to locate the information reported by a specific person.

I found the records for my grandparents along with those for my mother, who was born in December 1949. They lived in rural Lincoln County, Kentucky. My grandfather, Bernard Camenisch, born in Kentucky to a Swiss father, worked 92 hours the previous week as a dairy farmer. A decade earlier, in the 1940 Census, he was living with his father, also a farmer; he worked 60 hours the week prior to answering that census survey. My grandmother Dorothy was a "sample line respondent," and so answered questions asked to only one in five individuals.

My research program, with a range of coauthors, uses publicly available census data with names and other identifying information to create large panel datasets. This research follows men — who are easier than women to track from one census to the next — across decades in the US and other countries. These linked datasets enable us to answer a range of questions about the impact of early-life shocks on adult outcomes. For example, what was the effect of closing schools during the 1918 flu pandemic on children's later-life outcomes? What did the arrival in Southern counties of the boll weevil, a cotton-boll-eating beetle, do to children's school enrollment, and ultimately, educational attainment? What effect did the huge negative shock to family wealth of Emancipation have on the later-life economic standing of children of slave-holding families? How does migration feature in individual adjustments to environmental or immigration shocks?

Creating Linked Datasets

The digitization of the 1950 Census — its transformation from scanned images to a machine-readable database — is ongoing. This process was only completed in the past decade for US decadal censuses from 1850 through 1940. Researchers can access names, birthplaces, ages, occupations, and many other rich variables for every person enumerated in a specific census. Methods to link individuals across any combination of censuses rely on the fact that name, birth year, and birthplace do not change, for men at least, across decades.

Any linking method that uses these fixed characteristics to match observations across time faces some challenges. First, names are often spelled differently in different censuses by the time the data reaches researchers. The name could have been written incorrectly in the original source, the handwriting may be difficult to read, or there could be a basic transcription error. My grandfather's first name is listed as "Benard" in 1940. That is incorrect, but the handwriting is difficult to read. Second, not everyone remembers or knows their age. Particularly in a period when many people did not have birth certificates or had not gone to school for more than a few years, ages tend to be "heaped" — individuals are more likely to report mul-

Katherine Eriksson is a research associate in the NBER's Development of the American Economy Program and an associate professor of economics at the University of California, Davis. Prior to moving to UC-Davis, she was an assistant professor at California Polytechnic State University from 2013 to 2015. She serves on multiple editorial boards.

Eriksson's research interests focus on questions related to labor economics and demography in US history. Almost all of her papers use large-scale panel datasets created with linked datasets. She has worked extensively on immigration to the United States, as well as on questions in health, education, and incarceration.

Eriksson received a BS in mathematics and philosophy from Virginia Polytechnic Institute in 2004 and a BA in philosophy, politics, and economics from the University of Oxford in 2006. She holds an MS in applied and agricultural economics from Virginia Tech and received her PhD in economics from the University of California, Los Angeles in 2013.

She lives in California with her husband and their four dogs, including Cleopatra, her soul mate in the form of a Chihuahua, and Louisa, a ridiculously tiny yet opinionated terrier. She has completed three Ironman triathlons and is contemplating a fourth.

tiples of 10 and five. Lastly, sometimes there are multiple individuals with the same characteristics. John Smith, born in Alabama in 1855, will never be linkable because there are too many records with the same information.

Ran Abramitzky, Leah Platt Boustan, James Feigenbaum, Santiago Perez, and I evaluate various linking methods in the face of such challenges.[1] One way to address the first problem is to standardize names using a phonetic spelling algorithm. For example, "Eriksson" and "Eriksen" become the same phonetic name. This wouldn't help with my grandfather. Another is to calculate a Jaro-Winkler score, which measures how far apart two names are. In this case, the distance between "Benard" and "Bernard" is so small that it would likely count as close enough for a match.

To fix the problem of inaccurate dates of birth, researchers must trade off accuracy in the spelling of names with how close birth years are across sources. Sometimes we accept being a few years off if we are pretty sure they are the right person because of other characteristics. We use a range of methods that are more or less stringent in terms of how similar each variable has to be, as well as a range of strategies for assessing name similarity. Using what we call "ground truth" data — genealogical data similar to that for Bernard where we know the links are correct — we are able to assess the accuracy of each method. We consider two metrics: how often the algorithm actually picks up a match when it should, and how often it makes a correct match. Ideally one would maximize the first while minimizing the second, but there is a trade-off. Figure 1 shows what we consider to be the production possibility frontier; it gives researchers an idea of how to choose between methods when they value a larger sample or a smaller false positive rate.[2] Our research also considers the trade-offs between computation time and accuracy.

Alas, there are always caveats in linked data. First, linking techniques almost entirely focus on men since until recent times women usually changed their names upon marriage. My grandmother, Dorothy, is in none of my linked samples. Second, match rates, or the likelihood of linking a given man across datasets, are quite low — anywhere between 10 percent and 40 percent depending on the method, time horizon and time period, and population of interest. Match rates tend to be lower in the nineteenth century than in the twentieth. They are higher for White than for Black men. Lastly, linked samples are usually not entirely representative. Linked individuals are more likely to be of higher socioeconomic status because these individuals are more likely to report their names and ages correctly across time. Nonetheless, researchers can use sample sizes in the millions to achieve precise estimates, and our research provides recommendations on dealing with the drawbacks.

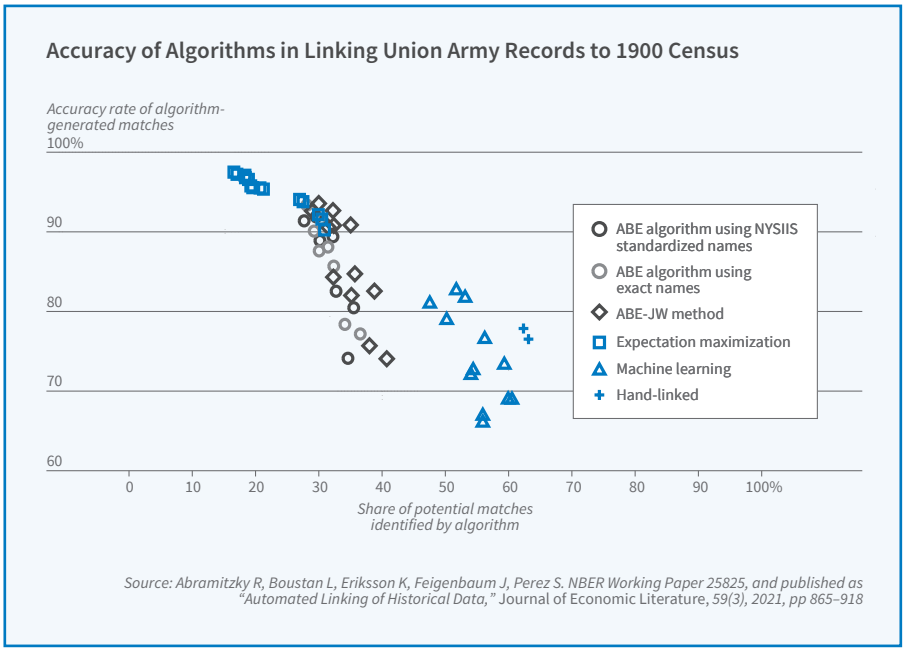

Figure 1

Connecting Childhood Shocks to Adult Outcomes

A range of recent research, including much of my own, has used these linked datasets to look at the effect of childhood shocks on adult outcomes. My most recent study, with Philipp Ager, Ezra Karger, Peter Nencka, and Melissa Thomasson, examines the effects of school closures during the 1918 flu pandemic.[3] Specifically, we ask if variation across cities in how long schools were shut in 1918–19 affected short-run school enrollment in 1920 as well as completed education by 1940, the first year the census asks about years of school. Why do we need linked data to do this? Why not just study a sample of men in the 1940 Census and assign school closures in childhood to them, without linking, based for example on where they live? If individuals selectively move, this approach results in measurement error and incorrect conclusions. We find no effect — a precise zero effect — of school closures on enrollment in 1920 and on educational attainment in the long run, likely because the school closures did not change behavior. Many schools were closed while the virus raged, and many students stayed home out of fear of an infectious virus.

In a recent paper, Richard Baker, John Blanchette, and I leverage linked data to examine the effect of the boll weevil, the beetle that crept across the South between 1895 and 1925, on children's completed education.[4] We argue that although the effect of this pest on school enrollment is theoretically ambiguous, the substitution effect — child labor became less productive as cotton productivity fell — dominated the income effect in this period, increasing school enrollment. Even though

there was limited access to high schools for Black students, we find that those who were at early ages when the boll weevil arrived completed almost 0.4 years of additional school.

Ager, Boustan, and I link childhood shocks to adult outcomes looking at the effect of the emancipation of slaves after the Civil War on the sons of men who owned slaves in 1860, on the eve of the war.[5] In this case, we must identify the exact wealth of fathers of sons whom we observe in censuses in 1880 and 1900, so linking is essential. Despite large hits to wealth, the sons of slaveowners were no worse off by 1900. We posit that the end of slavery had little effect on the relative position of groups in the South; sons of slaveowners still had their social networks and family connections even after losing a large amount of wealth.

Migration as an Adjustment Mechanism

Linked data allow us to study the mechanisms through which individuals adjust to economic and other shocks. Ager, Casper Worm Hansen, Lars Lønstrup, and I looked at the immediate and long-run effects of the San Francisco earthquake of 1906 on population and development in California, Nevada, and Oregon.[6] We found that the population of more-affected areas grew slower than that of less-affected areas in the six decades following the earthquake. This appears to be because new migrants to California chose less-affected places as their destinations.

In 1854, the Know-Nothing, or American, Party swept into power across the Northeast on the heels of the collapse of the Second Party System and the dissolution of the Whig Party in 1852. The party's platform varied across the country, but it was largely a nativist, anti-immigrant platform in the Northeast. Party leaders called for restriction of voting and naturalization rights of immigrants and deportation of "paupers and criminals." The party's supporters, heavily drawn from mid-skilled workers, blamed falling real wages on the influx of low-skilled Irish immigrants. However, this was also a time of rapid industrialization in the Northeast; mid-skilled jobs were quickly being replaced by low-skilled jobs as production moved into factories.

Marcella Alsan, Greg Niemesh, and I test these two competing claims about support for the Know-Nothings. Was it due to Irish labor market competition, or "de-skilling" due to industrialization?[7] Using yearly gubernatorial vote shares at the town level in Massachusetts, we construct indices of potential labor market competition imposed on natives by Irish immigrants and of the lowering of the skill content of jobs due to the movement of production from small shops into factories. We use the local share of employment-indifferent industries, as well as state-level shifts in industrial composition, along the lines of research by David Autor and coauthors.[8] We find that both variables positively predict Know-Nothing vote shares in the three years that the party won. To study how individuals in towns with more exposure to these shocks adjusted to them, we use linked data from 1850 to 1860 to track men who were more and less affected. We find that although wealth is lower in 1860 relative to 1850 for those affected by these shocks, this effect is somewhat tempered by migration. Leaving the county or state was more likely for those affected, and this led to smaller wealth losses.

The Future of Linked Data

Census linking has made great strides in the past decade due to newly available data and advances in computing technology. Hopefully, future work will allow women to be linked more successfully. Some studies are using marriage certificates, which include both birth and married surnames, to supplement census data. Researchers continue to access new data sources to enable linking across censuses and to add richness to the limited set of variables in the census data files. As more and more records become available from government and private sources, this linking can only grow.

---

1 "Automated Linking of Historical Data," Abramitzky R, Boustan L, Eriksson K, Feigenbaum J, Perez S. NBER Working Paper 25825, May 2019, and *Journal of Economic Literature* 59(3), September 2021, pp. 865–918. For another assessment of linking methods, see "How Well Do Automated Linking Methods Perform?" Bailey M, Cole C, Henderson M, Massey C. *Journal of Economic Literature* 58(4), December 2020, pp. 997–1044.
Return to Text

2 Linked US Census samples using some of these methods are available at CensusLinkingProject.org. More methods are being added over time.
Return to Text

3 "School Closures during the 1918 Flu Pandemic," Ager P, Eriksson K, Karger E, Nencka P, Thomasson M. NBER Working Paper 28246, December 2020.
Return to Text

4 "Long-Run Impacts of Agricultural Shocks on Education Attainment: Evidence from the Boll Weevil," Baker R, Blanchette R, Eriksson K. NBER Working Paper 25400, December 2018, and *Journal of Economic History* 80(1), March 2020, pp. 136–174.
Return to Text

5 "The Intergenerational Effects of a Large Wealth Shock: White Southerners after the Civil War," Ager P, Boustan L, Eriksson K. NBER Working Paper 25700, September 2019, and *American Economic Review* 111(11), November 2021, pp. 3767–3794.
Return to Text

6 "How the 1906 San Francisco Earthquake Shaped Economic Activity in the American West," Ager P, Eriksson K, Hansen C, Lønstrup L. NBER Working Paper 25727, April 2019, and *Explorations in Economic History* 77, July 2020.
Return to Text

7 "Understanding the Success of the Know-Nothing Party," Alsan M, Eriksson K, Niemesh G. NBER Working Paper 28078, November 2020.
Return to Text

8 For example, see "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure," Autor D, Dorn D, Hanson G, Majlesi K. NBER Working Paper 22637, December 2017. Forthcoming in *American Economic Review*.
Return to Text